Soft Computing (2019) 23:3037–3051
https://doi.org/10.1007/s00500-017-2951-6

METHODOLOGIES AND APPLICATION

Weighted network graph for interpersonal communication with temporal regularity

Ryoichi Shinkuma[1] · Yuki Sugimoto[1] · Yuichi Inagaki[1]

Published online: 25 November 2017

Abstract
Over the last decade, interpersonal communication has attracted more attention from researchers than before. Although the volume of data generated through various communication devices and tools could be enormous, the recent decrease in storage cost enables us to record and store it. The analysis of interpersonal communication is useful to estimate influence in social relationships among people, to detect communities, and to recommend potential friends for users on social networking services. A network graph, which is a mathematical model that represents people as nodes and past opportunities of interpersonal communication as edges, works in such analysis. However, when the capacity of the number of edges recordable in a graph database is limited, or when only a limited number of edges is used for high-speed analysis, it is still unclear which edges should be prioritized and utilized in the analysis. Previous studies suggested that edges in network graphs can be weighted on the basis of the aggregated duration of connections, the number of connections, or **the connection time**. However, temporal regularity in interpersonal communication has not been well considered in the previous studies. Therefore, in this paper, we propose an edge weighting method for network graphs from interpersonal communication that determines edge weighs on the basis of the scores obtained from the spectral analysis technique. The spectral analysis technique is utilized to numerically deal with temporal regularity and frequency of interpersonal communication. An examination using real records verifies that by using our edge weighting method, link prediction works better under a condition of the limited number of edges usable for the analysis. We also deeply analyze and present the distributions of the frequencies that characterize interpersonal communication.

Keywords Interpersonal communication · Weighted network graph · Temporal regularity · Frequency-domain analysis

1 Introduction

Over the last decade, interpersonal communication has attracted more attention from researchers than before. The first reason is the increase in the number of communication devices, such as personal computers, mobile phones, and smart phones. Second, communication tools have become diversified from traditional ones like telephone calls, e-mails, and short message services (SMSs) to message exchanges and photograph sharing on social networking services (SNSs). Furthermore, if wearable devices, which people wear all the time, become more popular in the future, face-to-face communication will be also observed as a kind of interpersonal communication (Hossmann et al. 2010). Although the volume of data generated through the above kinds of interpersonal communication could be enormous, the recent decrease in storage cost enables us to record and store it.

The analysis of interpersonal communication is useful to estimate influence in social relationships among people (Tang et al. 2012), to detect communities (Nguyen et al. 2011; Pandey et al. 2012; Rathnayaka et al. 2015), and to recommend potential friends for users on SNSs (Roth et al. 2010). A network graph, which is a mathematical model that represents people as nodes and interpersonal communication as edges, works in such analysis: it enables us to extract useful indicators from network graphs such as degree and centrality (Shinkuma et al. 2015). Traditionally, network graphs have been used for analyzing the World Wide Web (WWW) and SNSs (Scott 2000). Although in general edges in network

Communicated by V. Loia.

✉ Ryoichi Shinkuma
shinkuma@i.kyoto-u.ac.jp

[1] Kyoto University, Yoshida-honmachi, Sakyo-ku, Japan

graphs are manually registered like on the WWW and SNSs, in the analysis of interpersonal communication, edges are automatically established on the basis of past opportunities of interpersonal communication observed by the systems.

Graph databases (GDBs) are specialized for recording and storing the structural information of network graphs (Moniruzzaman and Hossain 2013). Although GDBs are beneficial thanks to their parallel-processing capability and scalability, (i) when the capacity of the number of edges recordable in a GDB is limited (Pokorný 2015), or (ii) when only a limited number of edges is used for high-speed analysis (Martínez-Bazan et al. 2011), GDBs do not give us any clear answer about which edges should be prioritized and utilized in the analysis. In this context, although traditionally an edge contains only information about its presence, edge weighting should be essential for efficient and effective use of GDBs. Edge weighting is also useful to visualize interpersonal communication using network graphs in order to highlight the characteristics of the relationships (Blagus et al. 2014). Edges formed from the observation in interpersonal communication could reflect more varied information like temporal characteristics of interpersonal communication. Previous studies suggested that edges in network graphs can be weighted on the basis of the aggregated duration of connections (Onnela et al. 2007), the number of connections (Wang et al. 2011), or **the connection time** (Kudelka et al. 2010; Hsu and Kshemkalyani 2015). However, temporal regularity in interpersonal communication has not been well considered in the previous studies. According to our intuition, people may make opportunities for interpersonal communication at the same time on the same day every week, and this kind of regularity should be prioritized in edge weighting more than ones between random pairs that occur randomly.

Therefore, in this paper, we propose a novel edge weighting method for network graphs, which determines edge weighs on the basis of the scores obtained from the spectral analysis technique. Although traditionally the spectral analysis technique has been widely used in communications engineering to analyze signals in the frequency domain, in our study, it is utilized to numerically deal with temporal regularity and frequency of interpersonal communication. Typically, if internal communication is periodic like exactly once a week, a peak in the spectrum at the 1-week frequency will be observed, while if it is random, the spectrum will be spread without any peak. Therefore, interpersonal communication of each pair could be scored on the basis of energy spectral density in the frequency domain. In our work, the examination using real records verifies that by using our edge weighting method, link prediction works better under a condition of the limited number of edges usable for the analysis. We also deeply analyze and present the distributions of frequencies that characterize interpersonal communication.

The rest of this paper is organized as below. Section 2 discusses the prior works related to our work. Section 3 proposes the system model and the edge weighting method of our system. Section 4 explains the evaluation scenario and the datasets for the validation of our system. Evaluation results are shown in Sect. 5, followed by the detailed analysis of frequencies that characterize interpersonal communication presented in Sect. 6. Finally, conclusions are drawn and future work is discussed in the last section.

2 Related works

2.1 Network graphs for interpersonal communication analysis

Network graphs have been used when researchers analyze interpersonal communication for various purposes like (i) to identify types of social relationships between people, (ii) to extract communities from a large group of people, and (iii) to assist people to find potential social relationships. **For (i),** Tang et al. (2012) formed network graphs from logs of phone calls, e-mails, Bluetooth scanning, and news sharing and used them to classifying types of social relationships; family, colleagues, or classmates. **For (ii),** Nguyen et al. (2011) worked an adaptive modularity-based method for identifying and tracing community structures in network graphs, which were formed from logs of e-mails and messages on SNSs. The information of identified community structures can be useful for developing not only social-aware strategies in SNSs but also efficient routing algorithms in mobile ad hoc networks (MANETs) and cellular networks. Pandey et al. (2012) have proposed and implemented a system that observes content publish/subscribe records of academic people and enables to automatically form communities for content sharing on the basis of the similarity of their interests. In the context of energy sharing in the smart-grid concept, Rathnayaka et al. (2015) have also proposed a framework that automatically forms virtual groups among energy prosumers using their past data of energy sharing. As a work related to (iii), Roth et al. (2010) discussed a friend-suggestion algorithm on the basis of network graphs constructed by the interactions between people and their groups thorough e-mails and other communication tools.

2.2 Edge weighting

As readers know, Granovetter is famous as the sociologist who originated a well-known hypothesis about weak ties: the proportional overlap of two individual's friendship networks varies directly with the strength of their tie to one another (Granovetter 1973). According to this hypothesis,

Table 1 Summary of prior and our works

Work	Index of edge weight	Meaning of edge weight
Onnela et al. (2007)	Aggregated call duration	Maintaining local communities/structural integrity
Wang et al. (2011)	The number of calls	Similarity in movements and connectedness
Kudelka et al. (2010)	Retention and stability with forgetting curve	Importance to characterize networks
Our work	Temporal regularity	Importance to characterize networks

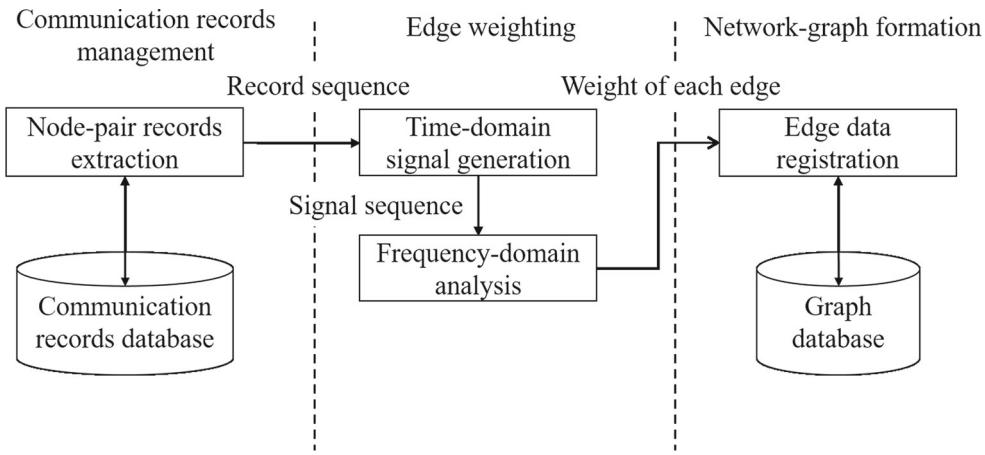

Fig. 1 System model

the strength of a tie between a pair of people could be determined by a combination of the amount of time, the emotional intensity, the intimacy, and the reciprocal services between the pair. However, it is still an open issue how to define the index of edge weights; a general idea is that the weight of an edge should be determined so as to reflect how big influence the edge gives to the social relationship.

Table 1 summarizes the prior works that tacked this issue. Onnela et al. used the aggregated duration of phone calls between people for edge weighting (Onnela et al. 2007). Their conclusion was that edges with low weights are crucial for maintaining the network's structural integrity, while edges with high weights play an important role in maintaining local communities. Wang et al. (2011) focused on how intense is the interaction between people by using the number of phone calls as edge weights. What they observed was that the similarity of individuals' movements and their social connectedness have strong correlation with the strength of interactions between them. Kudelka et al. (2010) introduced two parameters that represent temporal activity of edges: retention and stability. The forgetting curve, which is a well-known model obtained from the experiments about human memory, is used to calculate these parameters. They are effective to analyze the stability of edges and to reduce the network size based on that while maintaining the important components.

Our work considers temporal regularity of interpersonal communication as an index of edge weighting and tries to show how it works in the reduction of the size of network graphs, as shown in Table 1.

3 Proposed system

3.1 System model

We first describe the system model illustrated in Fig. 1. The proposed system is composed of the communication records management unit, the edge weighting unit, and the network-graph formation unit. The communication records management unit consists of the communication records database and the node-pair records extraction. It stores the logs of the interpersonal communication and inputs the data sequences to the edge weighting unit. The edge weighting unit includes the time-domain signal generation and the frequency-domain analysis. It puts weights of edges between nodes on the basis of the data sequences when interpersonal communication occurred and inputs the weights to the network-graph formation unit. The network-graph formation unit consists of the edge data registration and the GDB. It forms edges on the basis of the weights and registers the network graph in the GDB. Each function is described in detail below.

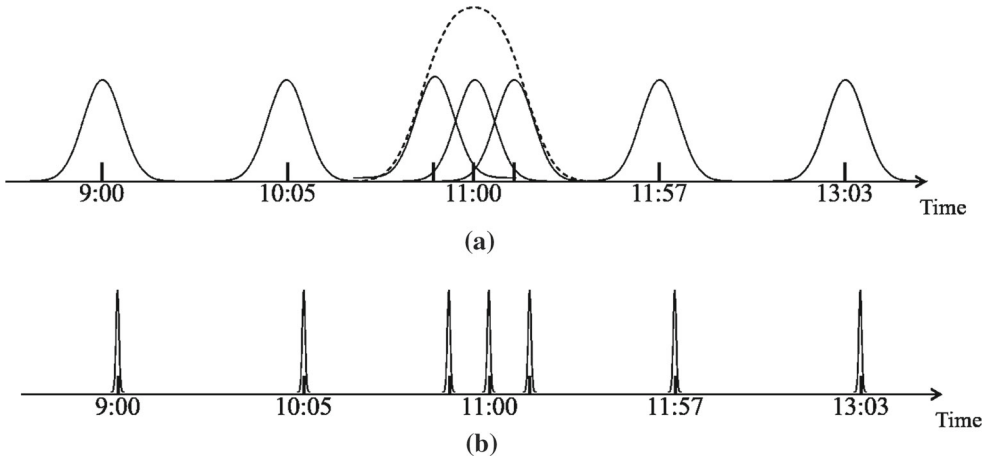

Fig. 2 Time function. **a** σ is large. **b** σ is small

3.2 Node-pair records extraction

The communication records database stores the logs of the interpersonal communication, which is observed in phone calls, SMS, e-mails, and message exchanges and photograph sharing on SNSs thorough the Internet or ad hoc networks by communication devices such as personal computers, mobile phones, and smartphones. Furthermore, wearable communication devices that people wear all the time are spreading, and face-to-face communication could be observed as a form of interpersonal communication (Hossmann et al. 2010). The node-pair records extraction generates the data sequences that represent the time of interpersonal communication between nodes by accessing the communication records database. Although the communication records database might hold information about the directions of interpersonal communication such as senders and receivers of e-mails or messages and the duration time of phone calls, such additional information is out of scope in this paper and will be considered in future work. If three or more nodes communicate at the same time through an opportunity of interpersonal communication like in e-mails addressed to two or more persons, the interpersonal communication is defined in all combinations of each node pair. Finally, the node-pair records extraction forwards the data sequences to the next block: time-domain signal generation.

3.3 Time-domain signal generation

The time-domain signal generation receives the data sequences received from the node-pair records extraction described in Sect. 3.2 and generates the time-domain signal. As shown in Fig. 2, the time-domain signal is generated by producing the time function when interpersonal communication occurred. The time function is given as:

$$g_{k_{i,j}}(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{-(t - T_{k_{i,j}})^2}{2\sigma^2}\right\} \tag{1}$$

where $g_{k_{i,j}}(t)$ is a Gaussian function and $k_{i,j}$ is the kth interpersonal communication between nodes i and j. $T_{k_{i,j}}$ is the time when $k_{i,j}$ occurred. The data sequences received from the node-pair records extraction have only the time information of interpersonal communication between nodes. Since those data sequences do not have amplitude components, it is impossible to capture their temporal characteristics even by analyzing them on the frequency domain. Therefore, we use the Gaussian function defined as Eq. (1). Considering extreme examples, when σ is close to infinity, which means that the time-domain signal is almost direct current, the frequency-domain signal has the component only when the frequency is zero. On the other hand, when σ is close to zero, the frequency-domain signal spreads infinitely to the ordinate direction. That is, since the Gaussian function works as a low-pass filter on the frequency domain, only the components of lower frequency are extracted as σ is large. For instance, when σ is set to 10 min, the Gaussian function spreads as shown in Fig. 2a. Then, the longer periods of interpersonal communication like 1 h or more on the frequency domain can be extracted, but the shorter periods cannot. For example, three interpersonal communications occurring around 11:00 could not be separated as shown in Fig. 2a. On the other hand, when σ is set to a smaller value, the Gaussian function becomes like that in Fig. 2b. In this example, since the period is deviated from 1 h by a few minutes, the deviations are observed as the high-frequency noise on the frequency domain. Note that t in Eq. (1) is a discrete value the unit time of which is Δt, and $g_{k_{i,j}}(t)$ is a discrete time signal.

The time functions given as Eq. (1) are summed up to obtain the time-domain signal. That is, signals overlapping at the same time are superimposed like the broken line in Fig. 2a. The time-domain signal $G_{i,j}(t)$ of the edge between nodes i and j is given as:

$$G_{i,j}(t) = \sum_{k=1}^{l} g_{k_{i,j}}(t) \tag{2}$$

where $l_{i,j}$ is the last interpersonal communication included in the data sequences. The time-domain signal generation forwards the time-domain signal with identifiers of nodes to the next step: frequency-domain analysis.

3.4 Frequency-domain analysis

The frequency-domain analysis converts the time-domain signal received from the time-domain signal generation described in Sect. 3.3 into the energy spectral density function by Fourier transform. The discrete Fourier transform for the time-domain signal of the edge between nodes i and j is given as:

$$F_{i,j}(f) = \sum_{t=T_{1_{i,j}}}^{T_{l_{i,j}}} G_{i,j}(t) \exp\left(-j\frac{2\pi f t}{N}\right) \tag{3}$$

where $N = (T_{l_{i,j}} - T_{1_{i,j}} + \Delta t)/\Delta t$. The energy spectral density function $\mathrm{ESD}_{i,j}(f)$ of the edge between nodes i and j is defined as:

$$\mathrm{ESD}_{i,j}(f) = |F_{i,j}(f)|^2 \tag{4}$$

The maximum values of the energy spectral density function of the edge between nodes i and j are detected, and the set of them is defined as $\mathbf{Peak_{i,j}}$. Note that the component the frequency of which is zero on the frequency domain, which means the direct current component, is ignored. On the basis of the hypothesis that the interpersonal communication that has regular characteristics in time should be prioritized over one that occurred randomly, the top n values in $\mathbf{Peak_{i,j}}$ are summed up and used as the weight. Where $\mathrm{rank}_{i,j}(m)$ is the mth largest value in $\mathbf{Peak_{i,j}}$, the weight of the edge between nodes i and j is given as:

$$W_{i,j} = \sum_{p=1}^{n} \mathrm{rank}_{i,j}(p) \tag{5}$$

Figure 3a–c shows the definition of the weight when $n = 1$, $n = 2$, and $n = 3$, respectively. Finally, the frequency-domain analysis forwards the weights with the identifiers of the nodes to the next step: edge data registration.

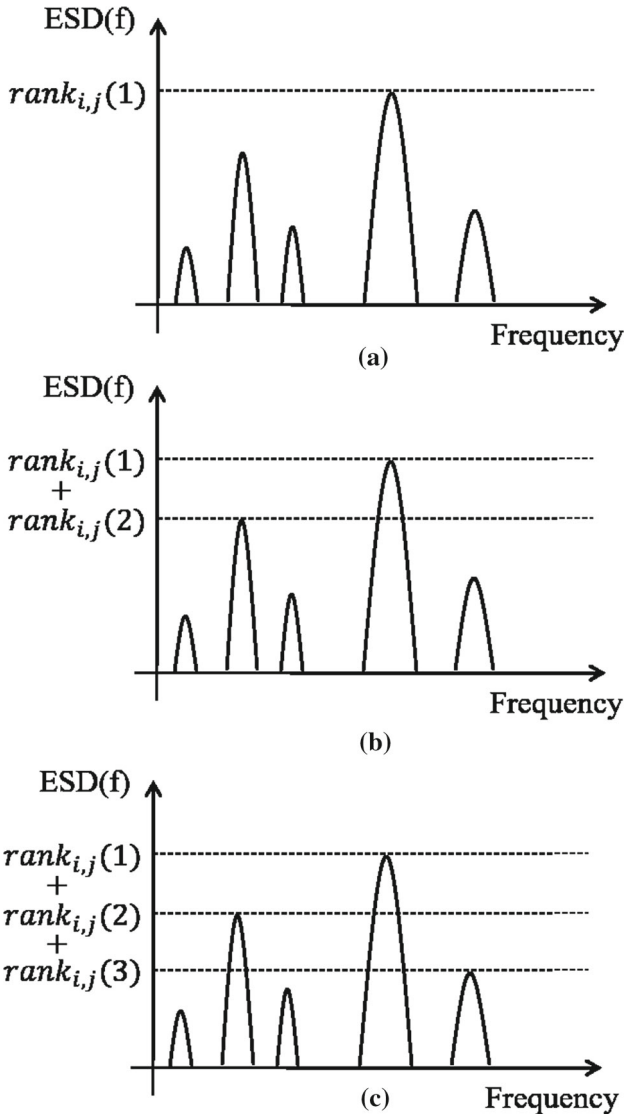

Fig. 3 Edge weighting. **a** $n = 1$. **b** $n = 2$. **c** $n = 3$

3.5 Edge data registration

The edge data registration registers the edges in the GDB on the basis of the weights obtained in the frequency-domain analysis described in Sect. 3.4. As mentioned in Sect. 1, the weights are utilized to thin out the edges from the original network graph (i) when the capacity of the number of edges recordable in a GDB is limited, or (ii) when only a limited number of edges is used for high-speed analysis. In the case of (i), the edges with higher weights are preferentially registered, and in the case of (ii), the edges are registered with the weights as additional information in the GDB.

4 Examination using real records

4.1 Evaluation scenario

We built an evaluation scenario to validate our edge weighting method. As we mentioned in the previous sections, we could say our method works well (i) when the capacity of the number of edges recordable in a GDB is limited, or (ii) when only a limited number of edges is used for high-speed analysis. Therefore, our evaluation scenario should be the one that can verify it. We adopted 'link prediction,' which is one of the most popular network-graph analyses, in our evaluation scenario. Link prediction is used to estimate unknown edges on the basis of the network structure composed of known edges and has been widely used for friend recommendation on SNSs (Sett et al. 2016). We produced network graphs for evaluation from datasets of interpersonal communication and applied link prediction to them. To assess how accurately link prediction performs, we masked 5, 10, and 20% of the edges randomly in advance from the original network graphs and estimated them as unknown edges from the rest of the (known) edges. Note that the masked edges had to be chosen from the ones between node pairs that had other two-hop or three-hop paths because, otherwise, it is impossible to estimate the masked edges by using link prediction. Furthermore, to consider the limitation of the number of edges usable in the analysis, we introduced a parameter r, which indicates the ratio of the number of usable edges to the total number. We varied r from 100 to 60% to observe how our method and the compared methods, which will be introduced shortly, performed.

In the link prediction process, we adopted Adamic and Adar (2003) and Katz (1953), which are ones of the most famous link prediction algorithms. In both methods, the relationships between every pair of nodes are scored. They estimate that an unknown edge between nodes A and B with the lager $E(A, B)$ exists more likely. $E(A, B)$ in Adamic/Adar is given as:

$$E(A, B) = \sum_{z \in |\Gamma(A) \bigcap \Gamma(B)|} \frac{1}{\log |\Gamma(z)|}, \tag{6}$$

where $\Gamma(x)$ is the set of neighbors of node x. On the other hand, $E(A, B)$ in Katz is given as:

$$E(A, B) = \sum_{l=1}^{\infty} \beta^l |\mathrm{paths}^{(l)}_{(A,B)}|, \tag{7}$$

where $|\mathrm{paths}^{(l)}_{(A,B)}|$ is the number of paths between nodes A and B with the hop length of which is l and β is a parameter between 0 and 1. While increasing l expands the range of how widely the algorithm considers known edges in the network graph for its scoring, decreasing β exponentially reduces the influence of paths with long hop lengths in the scoring. In our evaluation, l and β were set to 3 and 0.05, respectively.

4.2 Index for comparative evaluation

As described in Sect. 3, it is necessary to determine a couple of parameters in the proposed method. First, n, which means the first to nth largest values of energy spectral density are considered in scoring edge weights, were set to 1, 2, and 3. We believe it is reasonable not to consider a larger n than 3 because from our real-life experience, it should be unnatural that a certain pair of people have four or more regular period of their interpersonal communication like once an hour, once a day, once a week, and once a month. Second, the unit time Δt was set to 1 min, which is also reasonable because 1 min is detailed enough to consider temporal regularity of interpersonal communication. Third, the width of the Gaussian function σ was set to 10 min; from our real-life experience, when we say an interpersonal communication is 'temporally regular,' it should be acceptable that it happens in the 10-min range before and after the exactly determined time like the range of 11:50 to 12:10 as a temporally regular interpersonal communication at 12:00.

As benchmarks for our method, we introduced three methods: count-based, period-based, and random method. The count-based method determines edge weights in proportion to the number of opportunities in each interpersonal communication during a given period. The period-based method determines edge weights in proportion to the time span from the first to last opportunities in each interpersonal communication. Note that in the period-based method, if the number of interpersonal communication opportunities between a pair of nodes is only one, the weight of the edge between them is set to zero. The random method determines weights of edges randomly, which was introduced as the method that gives the lower-bound performance. In these conventional methods, like the proposed method, the top r% of the edges with larger weights were utilized for link prediction.

We adopted the area under the curve (AUC) value, which has been widely used for evaluating the performance of link prediction (Lichtenwalter et al. 2010; Xu-Rui et al. 2015), as the index for the comparative evaluation. The AUC value is the area under the curve of the receiver operating characteristic (ROC) and ranges from 0 to 1. The curve of the ROC is plotted with the false-positive rate (FPR) values on the abscissa and the true-positive rate (TPR) values on the ordinate that are computed with many different thresholds. Table 2 shows the confusion matrix, which explains the definition of true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Then, FPR and TPR are defined as:

Table 2 Confusion matrix

Predicted\actual	Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

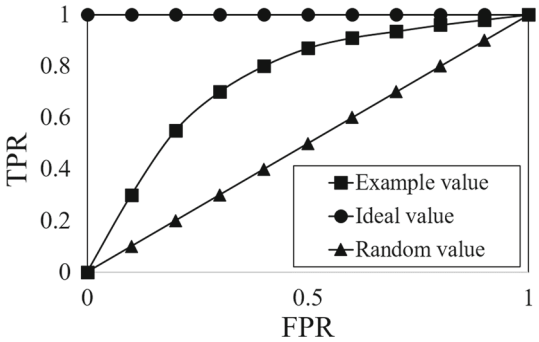

Fig. 4 Example of the curve of the ROC

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{8}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

Figure 4 shows an example of the curve of the ROC. In the ideal case (complete prediction) and in the worst case (random prediction), the AUC becomes 1.0 and 0.5, respectively.

4.3 Datasets

Table 3 summarizes the three datasets we used in our evaluation: Enron, Irvine, and Friends and Family. All of them have been well used in the research fields of social networks and communication networks. The details of each dataset are described as below.

The Enron dataset is composed of over 600,000 e-mails sent and received by 158 employees of the Enron Corporation. Those e-mails were acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse (Klimt and Yang 2004). In our evaluation, we used the data of the period from January 1, 2000, until April 30, 2000, which includes 3728 people and 49,889 e-mails after eliminating duplicate and unsent e-mails.

The Irvine dataset is composed of 59,835 messages exchanged by 1899 users of the Facebook-like SNS that originated from an online community for students at University of California, Irvine. The identifiers of the people and the time when the messages were sent and received are open for public (Opsahl and Panzarasa 2009). We used the data of the period from May 1, 2004, until August 31, 2004, which includes 1326 people and 51,728 messages.

The Friends-and-Family dataset was collected by the experiment of Massachusetts Institute of Technology (MIT).

Table 3 Three real datasets

Dataset	Type	Period (from/until)	Node	edge
Enron	E-mail	Jan. 1 2000/Apr. 30 2000	3728	7526
Irvine	SNS	May 1 2004/Aug. 31 2004	1326	7564
Friends and Family	SMS	Jan. 1 2011/Apr. 30 2011	1130	1344

Table 4 Number of masked edges

Dataset	5%	10%	20%
Enron	376	752	1505
Irvine	378	756	1512
Friends and Family	67	132	–

We chose SMS logs for our evaluation though the dataset also includes logs of phone calls, logs of Bluetooth connections, and answers for a questionnaire survey (Aharony et al. 2011). The subjects were members of a young-family residential living community adjacent to MIT. The pilot phase with 55 participants was launched in March 2010. After that, 130 participants from approximately 64 families participated in the second phase of the study, which started on September 2010. Including logs of messages sent and received by other people than these participants, the dataset contains 1130 people and 35,232 messages.

All the three datasets described above include the time stamps that indicate when interpersonal communication opportunities between each pair of people occurred. We here have to consider that it could be different what time a message is sent at from what time it is received at. In our evaluation, if the difference is larger than 1 min, we dealt with them separately as two interpersonal communication opportunities. However, if two or more messages are exchanged between a pair of people within 1 min, since we set the unit time to 1 min, we considered them as a single opportunity. As summarized in Table 3, the numbers of the edges of the network graphs formed from Enron, Irvine, and Friends and Family were 7526, 7564, and 1344, respectively. As described in Sect. 4.1, 5, 10, and 20% of the edges of the original network graphs were masked at random in advance as unknown edges. Table 4 shows the number of the masked edges for network graphs formed from each dataset. However, the 20% case cannot be examined in the Friends-and-Family dataset. This is because, as described in Sect. 4.1, ones between node pairs that have at least a 2-hop or 3-hop path can be masked; the ratio of node pairs that satisfied this condition was smaller than 20% in the Friends-and-Family dataset.

5 Results of AUC value

This section shows the results obtained from our examination and discusses the following five points: (1) the most appro-

priate value of n in Eq. (5) in the proposed method, (2) the superiority of the proposed method to the benchmark methods, (3) how the performance varies for different datasets, (4) how the performance varies for different link prediction methods, and (5) how the performance varies for the ratio of the number of masked edges to the total one.

5.1 5% masked case

In this section, we discuss the results when 5 % of the edges was masked as unknown edges. Figures 5 and 6 show the results using Adamic/Adar and Katz, respectively. In each figure, (a), (b), and (c) show the results of the Enron, Irvine, and Friends-and-Family datasets, respectively. First, let us look at the result in Fig. 5a. The abscissa of the figure is the AUC value, while the ordinate is r, which is the ratio of the number of the usable edges to the total one in the network graphs, as described in the previous section. When $r = 1$, since all the edges in the network graphs are used, there is no difference among all the methods. On the other hand, when $r = 0.6$, the top 60 % of the edges with the higher weights are used for link prediction. In the figure, the results of the proposed method ($n = 1, 2, 3$), the count-based method, the period-based method, and the random method are plotted. As we see in Fig. 5a, it was found that as r was smaller, the AUC values became smaller. This is simply because link prediction became more difficult when the number of usable edges was limited. As we expected, the random method gave the lowest performance among all the methods, which demonstrated the effectiveness of edge weighting on the basis of temporal characteristics of interpersonal communication. As we see in Fig. 5a, the proposed method was superior to the other methods. Next, in the comparison of the proposed methods with n =1, 2, and 3, by setting n to 1, the result was the best when r was between 0.7 and 0.9. In other words, taking the period giving the maximum value of the energy spectral density into account was effective enough rather than considering the periods giving the second and third maximum values when usable edges were limited to 70–90 %, which is a realistic and reasonable range (i) when the capacity of the number of edges recordable in a GDB is limited, or (ii) when only a limited number of edges is used for high-speed analysis. However, the cases of $n = 2$ and 3 were slightly better than $n = 1$ when r was 0.6 and 0.65. This observation suggests, as the number of unusable edges is more strictly limited, considering the periods giving the second and third maximum values of the energy spectral density becomes more meaningful to maintain the prediction accuracy.

As shown in Fig. 5b, c, the results obtained using the other datasets also presented the three trends we have already observed: the random method gave the lowest performance; the proposed method was the most effective of all; the

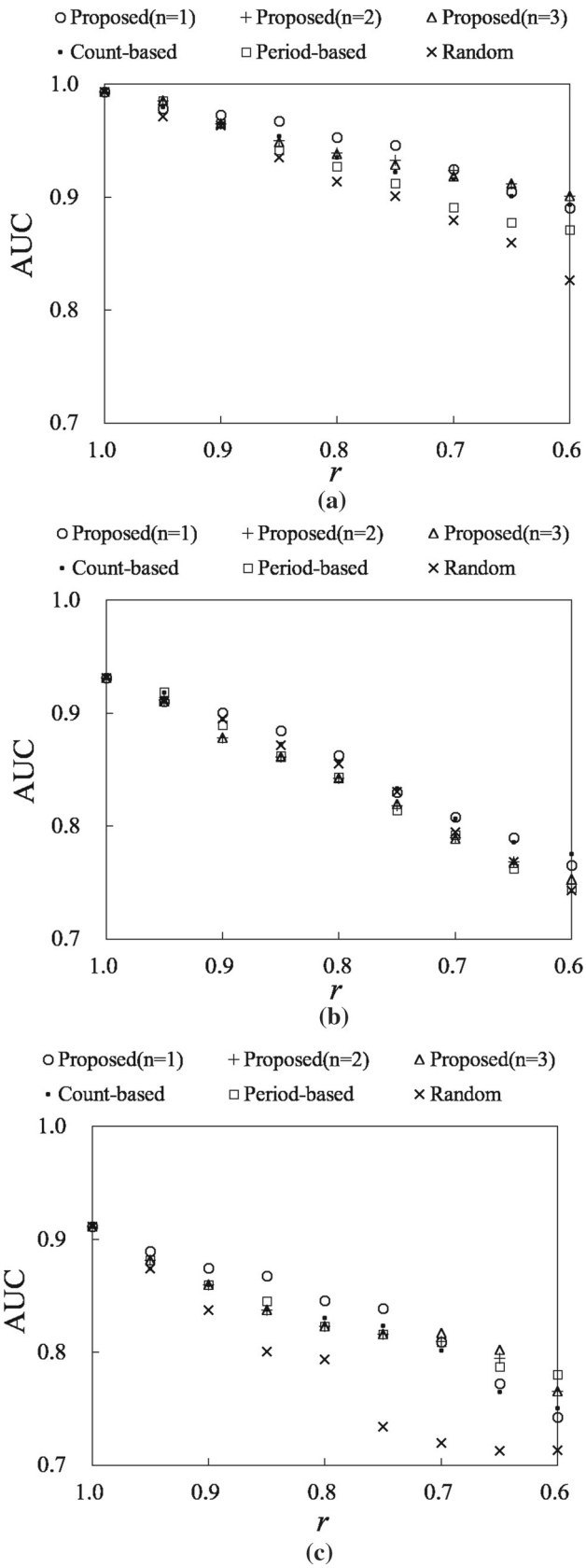

Fig. 5 AUC result. Adamic/Adar was used. 5% of edges were masked. **a** Enron. **b** Irvine. **c** Friends and Family

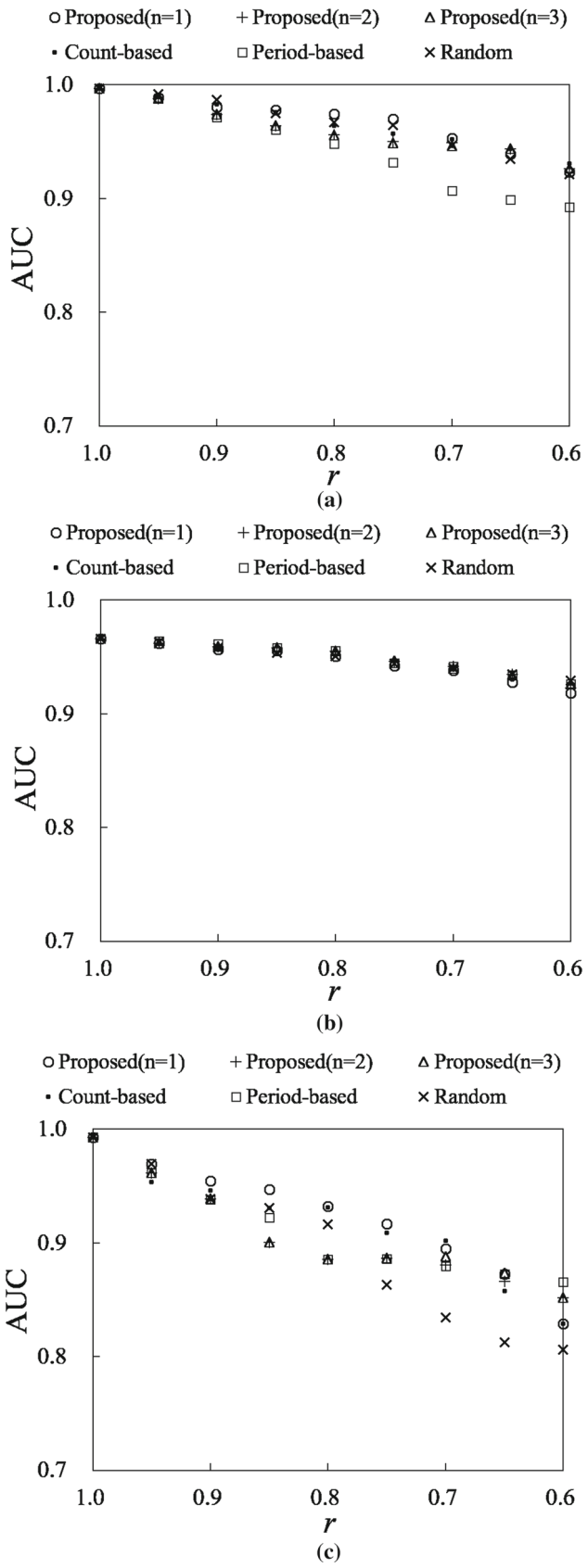

Fig. 6 AUC result. Katz was used. 5% of edges were masked. **a** Enron. **b** Irvine. **c** Friends and Family

proposed method when $n = 1$ was basically better than when $n = 2$ and 3. However, we see a couple of different observations dataset by dataset. For example, in Fig. 5b, the difference between the random method and other methods was small. Another example is that, in Fig. 5c, setting $n = 2, 3$ in the proposed method was the most effective of all the methods when r was 0.7 or smaller. From the discussion about the results in Fig. 5, we have reached the following conclusions: (1) In the proposed method, it is the most effective in the realistic range of usable edges to use only the maximum value of the energy spectral density for edge weighting by setting $n = 1$; (2) the proposed method performs best as long as n is set appropriately; (3) the proposed method works well for various types of datasets.

Next, we discuss Fig. 6a–c. Compared with Fig. 5, the AUC values were basically higher in Fig. 6. This is reasonable because Katz was developed to improve the link prediction performance against the classical methods like Adamic/Adar, which only considers 2-hop relationships. We confirmed the following three trends we had observed in Fig. 5: the random method gave the lowest performance; the proposed method was the most effective of all; in the proposed method, $n = 1$ was better than $n = 2$ and 3. However, the results were slightly different among different datasets. The period-based method was less effective than the random method in Fig. 6a. The difference among different methods was small in Fig. 6b. Fortunately, the overall observation is the same as the one from Fig. 5: (1) In the proposed method, it is the most effective in the realistic range of usable edges to use only the maximum value of the energy spectral density for edge weighting by setting $n = 1$; (2) the proposed method performs best as long as n is set appropriately; (3) the proposed method works well for various types of datasets. In addition, through the discussion from Figs. 5 and 6, it has been verified that (4) the proposed method performs for different link prediction methods.

5.2 10 and 20% masked cases

In this section, we see how link prediction performs when 10 and 20 % of the edges were masked as unknown edges. We first show how Adamic/Adar and Katz performed when 10 % was masked in Figs. 7 and 8, respectively. (a), (b), and (c) show the results of the Enron, Irvine, and Friends-and-Family datasets, respectively. Basically, the trends were very similar to the ones we observed for the case of 5 % in the previous section. That is, points of (1) to (4) described in the previous section were also applicable for the case of 10 %.

Next, we show how Adamic/Adar and Katz performed when 20 % is masked in Figs. 9 and 10, respectively. (a) and (b) show the results of the Enron and Irvine datasets, respectively. The result with the Friends-and-Family dataset is not available because, as mentioned in Sect. 4, 20 % of the

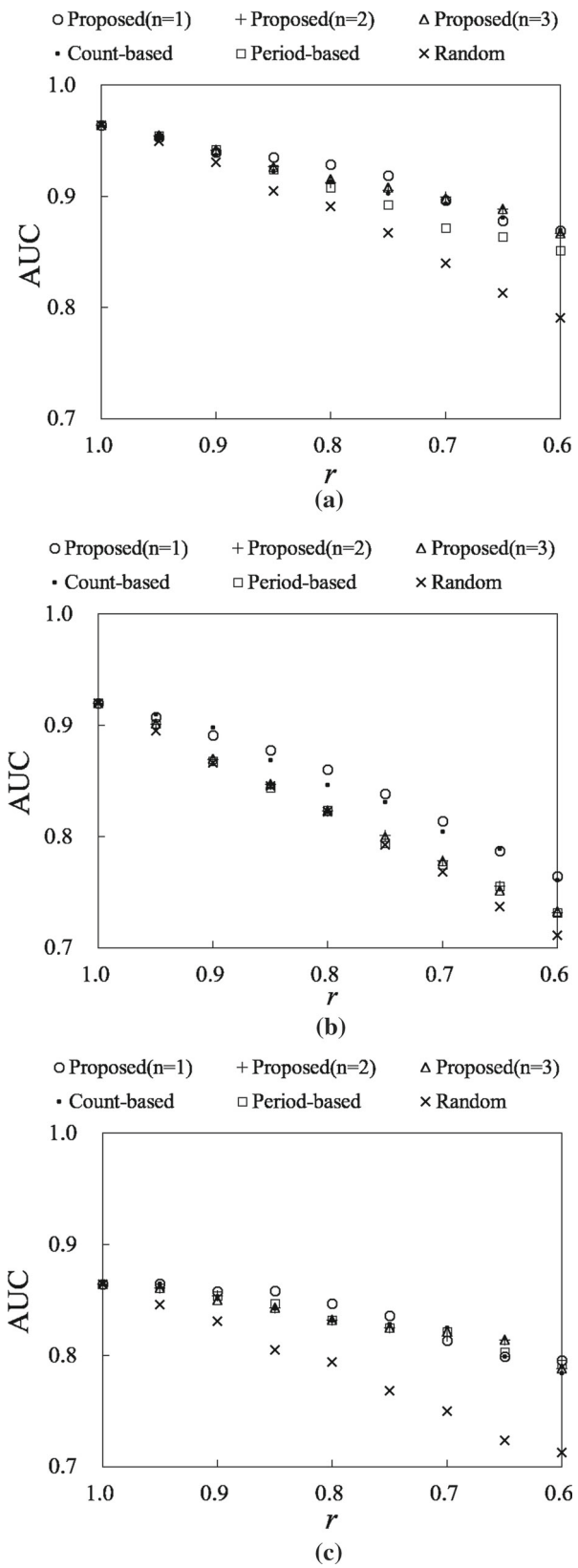

Fig. 7 AUC result. Adamic/Adar was used. 10% of edges were masked. **a** Enron. **b** Irvine. **c** Friends and Family

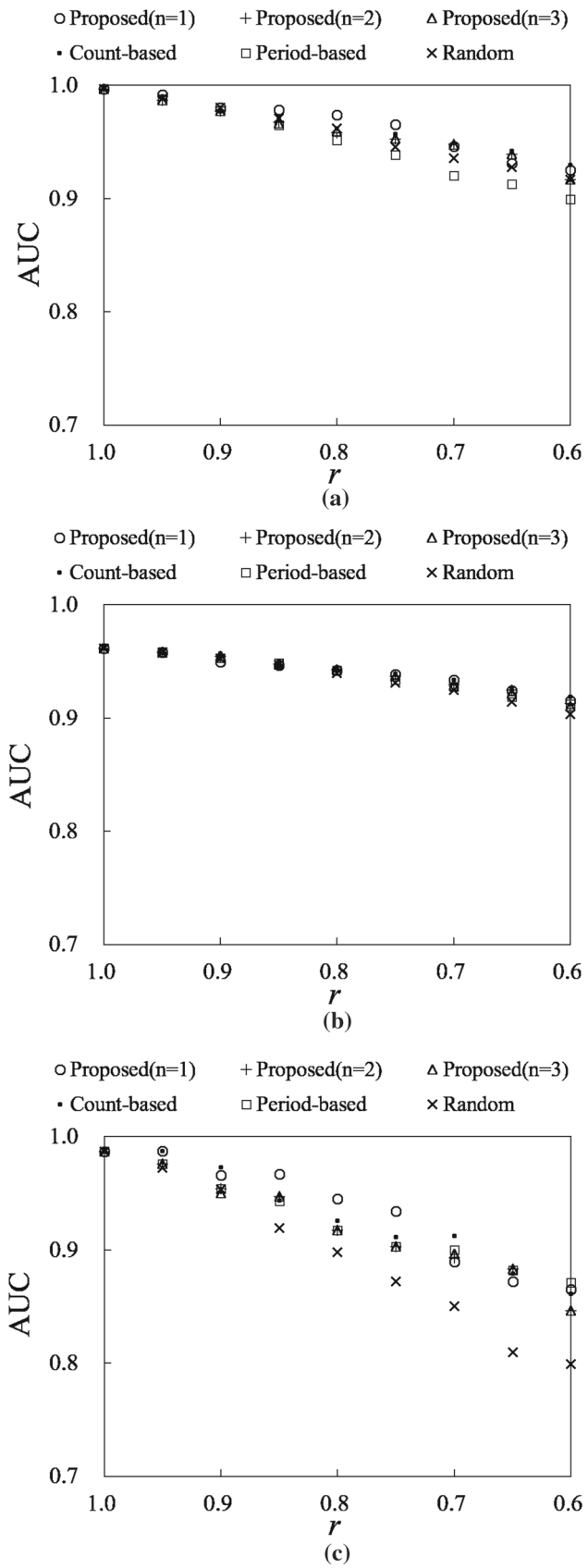

Fig. 8 AUC result. Katz was used. 10% of edges were masked. **a** Enron. **b** Irvine. **c** Friends and Family

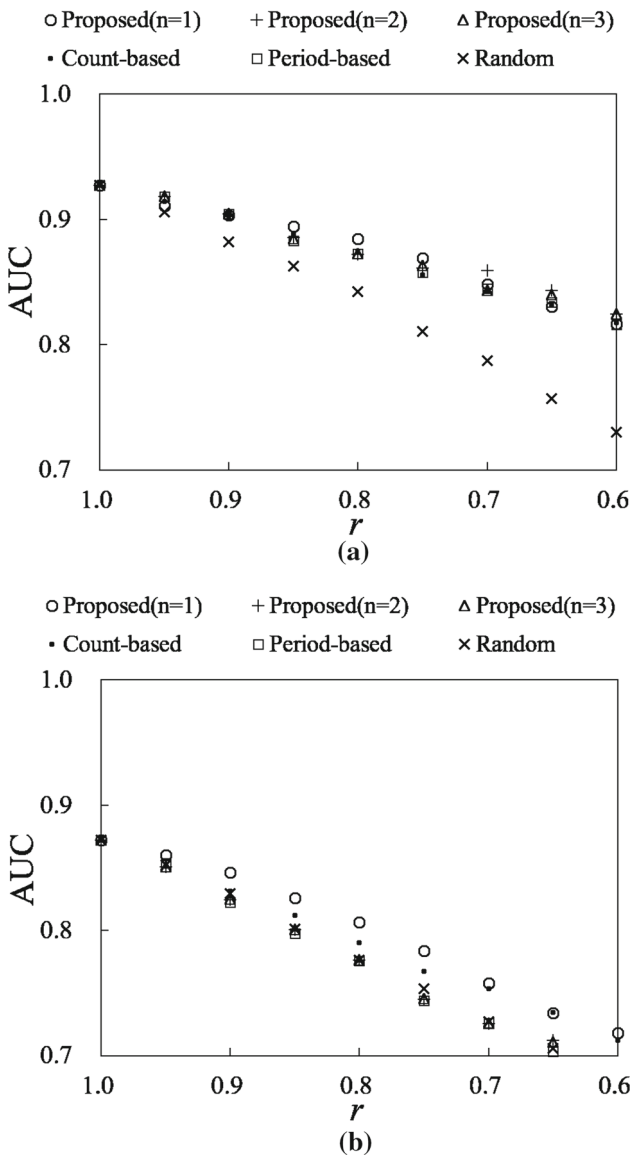

Fig. 9 AUC result. Adamic/Adar was used. 20% of edges were masked. **a** Enron. **b** Irvine

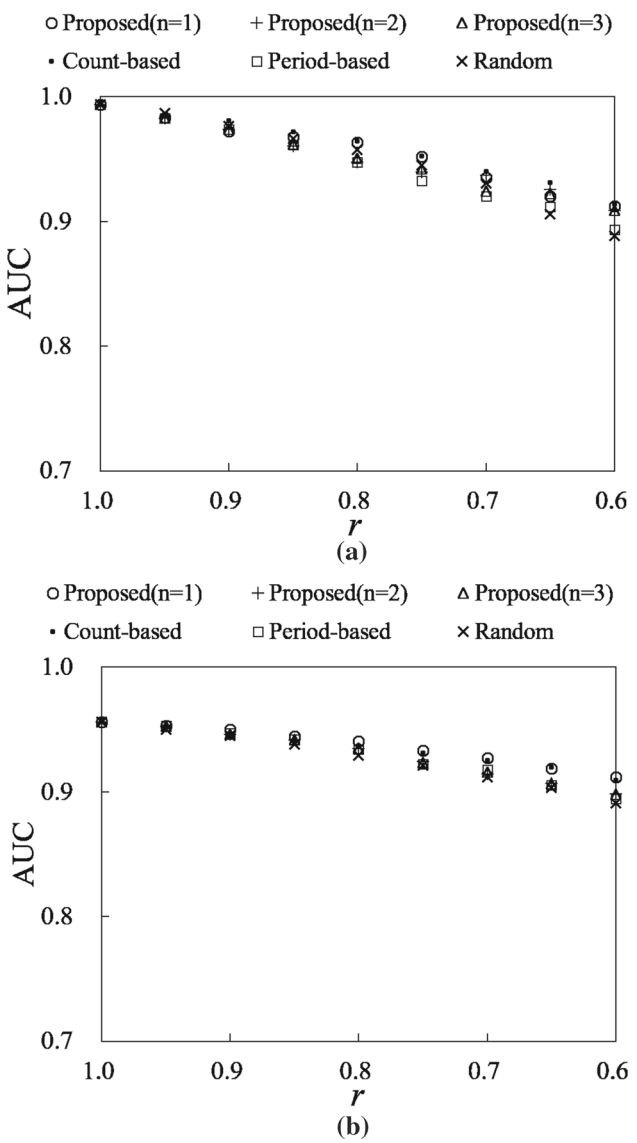

Fig. 10 AUC result. Katz was used. 20% of edges were masked. **a** Enron. **b** Irvine

edges could not be masked from its network graph. When we used Adamic/Adar, the AUC values in Fig. 9 were basically lower than those of Figs. 5 and 7. This is reasonable simply because link prediction became more difficult as the numbers of known and unknown edges decreased and increased. However, it was found that, when we used Katz, the AUC values in Fig. 10 did not decrease compared with those in Figs. 6 and 8 even if the number of unknown edges increased to 20%. This suggests that Katz is more tolerant against the decrease in known edges than Adamic/Adar thanks to its prediction capability. While the period-based method is the worst in Figs. 6a and 8a, there was no difference between the period-based method and other methods in Fig. 10a. Points (1) to (4) described above are also true in Figs. 9 and 10. In addition, through the discussion from the previous to this section, it has been also verified that (5) the proposed method works robustly against the decrease in known edges.

6 Distributions of frequencies in interpersonal communication

In this section, we discuss frequencies that characterize interpersonal communication used for edge weighting in the proposed method. As defined in Eq. (5), the proposed method uses the frequencies with the first to nth largest energy spectrum densities according to n (= 1, 2, or 3).

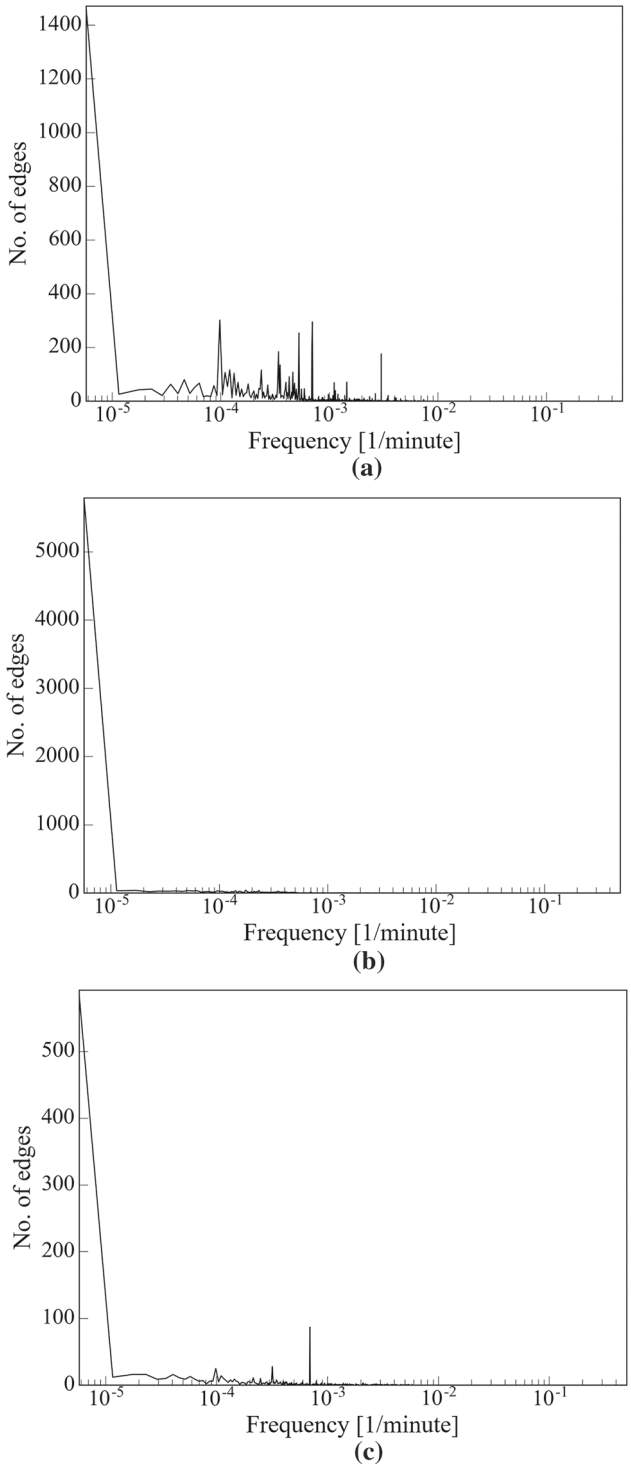

Fig. 11 Distribution of frequencies giving the first maximum value. **a** Enronl. **b** Irvine. **c** Friends and Family

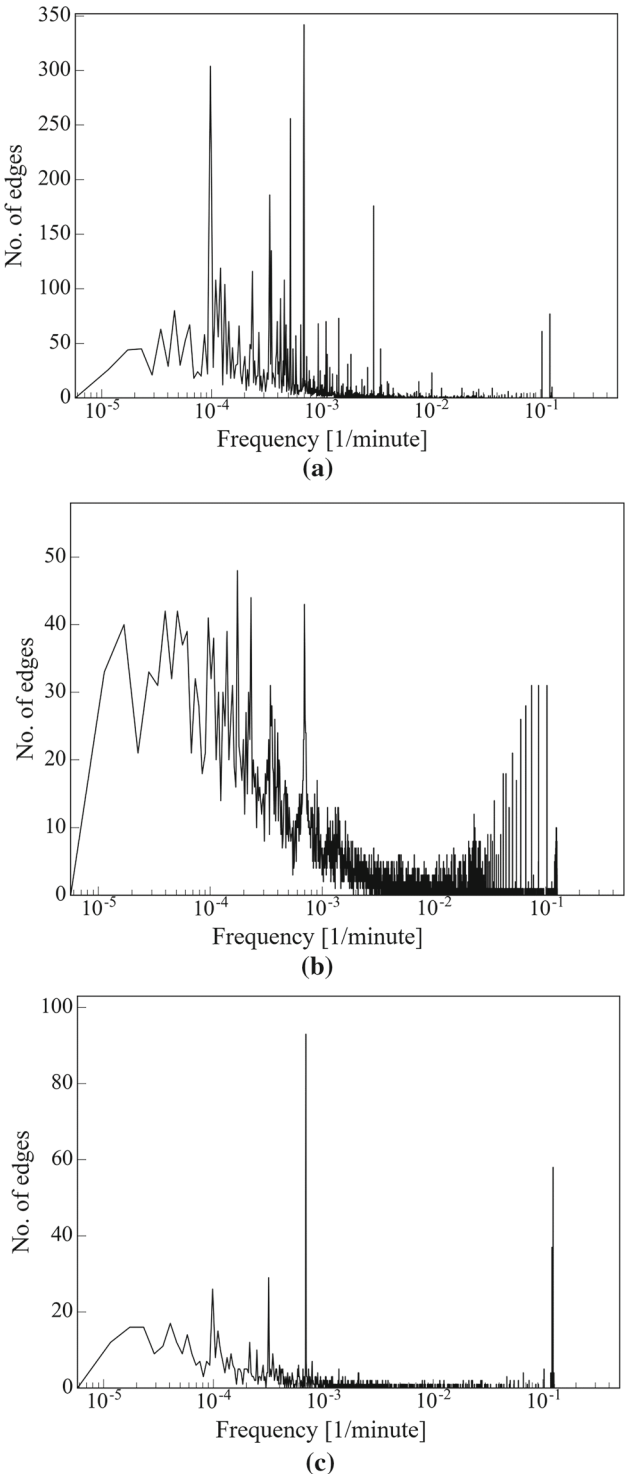

Fig. 12 Distribution of frequencies giving the second maximum value. **a** Enron. **b** Irvine. **c** Friends and Family

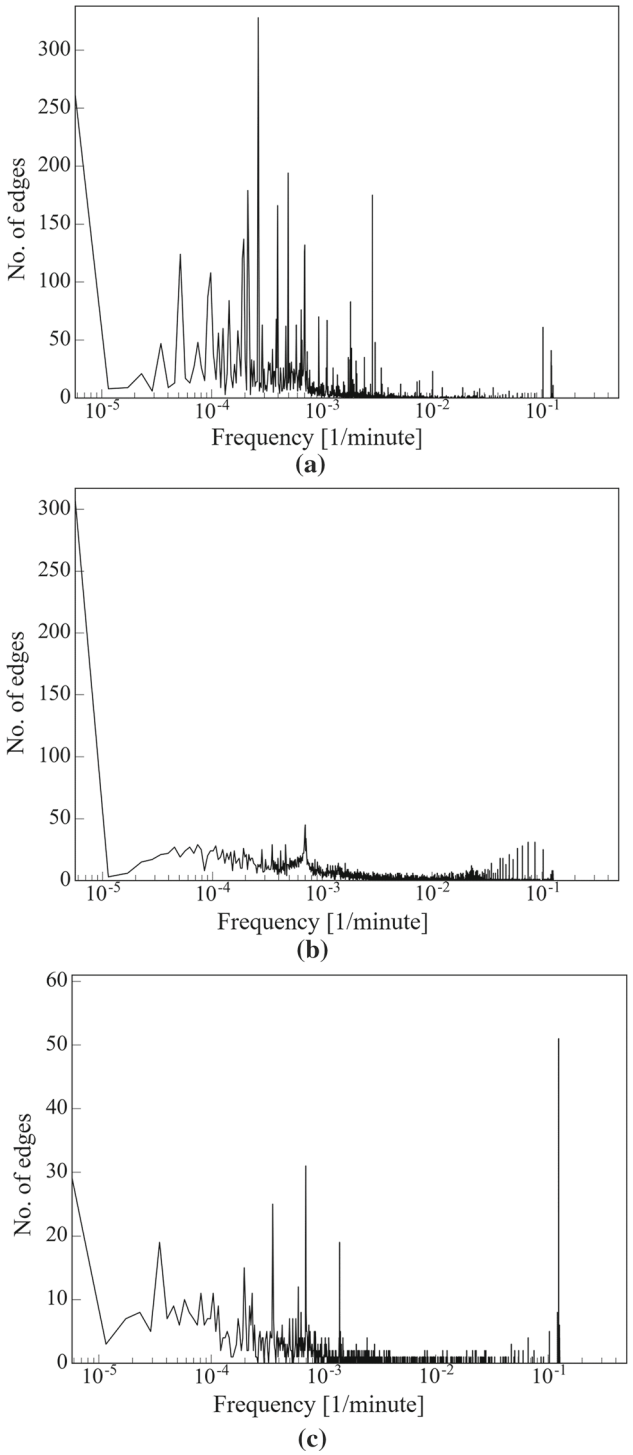

Fig. 13 Distribution of frequencies giving the third maximum value. **a** Enron. **b** Irvine. **c** Friends and Family

6.1 Frequencies giving first maximum value of energy spectral density

Figure 11 shows the distribution of the frequencies giving the maximum values of the energy spectral density in each edge. (a), (b), and (c) show the results of the Enron, Irvine, and Friends-and-Family datasets, respectively. Note that the ordinate is scaled according to the range of plots in each figure. What we first see in these figures is a concentration of spectrum at around 1.0×10^{-5} [/min]. This just came from the fact that the span of each dataset is 4 months as shown in Table 3; if the number of interpersonal communication between a certain pair in that span is only twice, the frequency becomes close to 1.0×10^{-5} (1/86000) [/min]. Next, in the result of the Enron dataset shown in Fig. 11a, another concentration is seen at around 1.0×10^{-4}. Making interpersonal communication opportunities once a week is equivalent to the frequency close to 1.0×10^{-4} (1/10080) [/min], which is consistent with our real-life experience that people at work tend to send and receive e-mails once a week for weekly meetings or reports. Another concentration of spectrum is observed at around 5.0 and 7.0×10^{-4}, which is equivalent to once a day. The concentration of spectrum observed at around 3.0×10^{-3} implies that some pairs of people made interpersonal communication once every 6 h. Thus, the above result from the frequency analysis suggests that interpersonal communication extracted from the Enron dataset was regularly done at frequencies, which were consistent with our intuition, and our method worked based on the frequencies as we had expected.

Next, we discuss the results obtained using the Irvine and Friends-and-Family datasets, which are shown in Fig. 11b, c, respectively. First, in Fig. 11b, there is no noticeable concentration of spectrum except around 1.0×10^{-5}. This means that interpersonal communication extracted from the Irvine dataset does not have frequencies that characterize it, which explains even different methods gave similar performance in the results obtained using the Irvine dataset as shown in Sect. 5. On the other hand, in the Friends-and-Family dataset, a clear concentration was observed at around 7.0×10^{-4}. This corresponds to once a day (1/1440 [/min]), which is a reasonable frequency as the one for sending and receiving SMS messages among friends or family. Such clear concentration is helpful for our method to prioritize edges based on their temporal regularity, which explains the reason why our method worked very well in the evaluation using the Friends-and-Family dataset as shown in Sect. 5.

6.2 Frequencies giving the second and third largest values of energy spectral density

Figures 12 and 13 show the frequencies giving the second and third maximum values of the energy spectral density function, respectively. In both figures, (a), (b), and (c) present the results obtained using the Enron, Irvine, Friends-and-Family datasets, respectively. In the results of the Enron dataset shown in Figs. 12a and 13a, we see a concentration of spectrum at around 1.0×10^{-4} as we saw in Fig. 11a. In the results

of the Irvine dataset shown in Figs. 12b and 13b, there is no clear concentration at any specific frequency. In the results of the Friends-and-Family dataset shown in Fig. 12c, surprisingly, we see a concentration at the frequency equivalent to the period of 10 min; quite close interpersonal communication among friends and family was observed. Thus, through the overall observation from Sect. 6.1 to this section, by analyzing frequencies giving the maximum values of the energy spectral density function, it has been clarified that the proposed method appropriately used frequencies that characterize interpersonal communication for weighting edges in the network graphs.

7 Conclusions and future work

As for edge weighting, temporal regularity in interpersonal communication has not been well considered in the previous studies. Therefore, we proposed an edge weighting method for network graphs that determines edge weighs on the basis of the scores obtained from the spectral analysis technique. By the spectral analysis technique, interpersonal communication of each pair could be scored on the basis of energy spectral density in the frequency domain. The examination using real records verified that by using our edge weighting method, link prediction works better under a condition of the limited number of edges usable for the analysis. We also showed that our method works robustly against different datasets, a variety of link prediction methods, and the decrease in known edges. Furthermore, we presented the distributions of the frequencies of interpersonal communication in each dataset, which clarified that the proposed method appropriately used frequencies that characterize interpersonal communication for its weighting edges in the network graphs. Future work will be to consider the directions of interpersonal communication like sender and receiver in e-mail communication and to examine a mixed index of temporal regularity and temporal duration of interpersonal communication. Another remaining issue is to examine the computational complexity and the scalability of our method, which would be improved by distributed approaches like the one proposed in Steinbauer and Kotsis (2016).

Acknowledgements This study was funded by KDDI foundation, Japan.

Compliance with ethical standards

Conflict of interest Ryoichi Shinkuma has received research grants from KDDI foundation, Japan. Yuki Sugimoto declares that he has no conflict of interest. Yuichi Inagaki declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

Adamic LA, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230

Aharony N, Pan W, Ip C, Khayal I, Pentland A (2011) SocialfMRI: investigating and shaping social mechanisms in the real world. Pervasive Mob Comput 7(6):643–659

Blagus N, Subelj L, Bajec M (2014) Assessing the effectiveness of real-world network simplification. Physica A Stat Mech Appl 413:134–146

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Hossmann T, Spyropoulos T, Legendre F (2010) Know thy neighbor: towards optimal mapping of contacts to social graphs for DTN routing. In: Proceedings of the IEEE INFOCOM, IEEE, pp 1–9

Hsu T-Y, Kshemkalyani AD (2015) Variable social vector clocks for exploring user interactions in social communication networks. Int J Space Based Situated Comput 5(1):39–52

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–44

Klimt B, Yang Y (2004) The enron corpus: a new dataset for email classification research. In: European conference on machine learning, Springer, pp 217–226

Kudelka M, Horak Z, Snasel V, Abraham A (2010) Social network reduction based on stability. In: 2010 international conference on computational aspects of social networks, IEEE, pp 26–28

Lichtenwalter RN, Lussier JT, Chawla NV (2010) New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 243–252

Martínez-Bazan N, Gómez-Villamor S, Escalé-Claveras F (2011) DEX: a high-performance graph database management system. In: IEEE 27th international conference on data engineering workshops, IEEE

Moniruzzaman ABM, Hossain SA (2013) NoSQL database: new era of databases for big data analytics–classification, characteristics and comparison. Int J Database Theory Appl 6(4):1–14

Nguyen NP, Dinh TN, Xuan Y, Thai MT (2011) Adaptive algorithms for detecting community structure in dynamic social networks. In: Proceedings of the IEEE INFOCOM, IEEE

Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, kaski K, Kertész J, Barabási A-L (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci 104(18):7332–7336

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Netw 31(2):155–163

Pandey M, Pathak VK, Chaudhary BD (2012) A framework for interest-based community evolution and sharing of latent knowledge. Int J Grid Util Comput 3(2/3):200–213

Pokorný J (2015) Graph databases: their power and limitations. In: IFIP international conference on computer information systems and industrial management, Springer, pp 58–69

Rathnayaka AJD, Potdar VM, Dillon TS, Kuruppu S (2015) Formation of virtual community groups to manage prosumers in smart grids. Int J Grid Util Comput 6(1):47–56

Roth M, Ben-david A, Deutscher D, Flysher G, Horn I, Leichtberg A, Leiser N, Matias Y, Merom R (2010) Suggesting friends using the implicit social graph. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 233–242

Scott J (2000) Social network analysis. SAGE Publications, London

Sett N, Sanasam SR, Sukumar N (2016) Influence of edge weight on node proximity based link prediction methods: an empirical analysis. Neurocomputing 172:71–83

Shinkuma R, Sawada Y, Omori Y, Yamaguchi K, Kasai H, Takahashi T (2015) A socialized system for enabling the extraction of potential values from natural and social sensing. In: Modeling and processing for next-generation big-data technologies, vol 4. Springer, pp 385–404

Steinbauer M, Kotsis GA (2016) DynamoGraph: extending the Pregel paradigm for large-scale temporal graph processing. Int J Grid Util Comput 7(2):141–151

Tang J, Lou T, Kleinberg J (2012) Inferring social ties across heterogenous networks. In: Proceedings of the fifth ACM international conference on Web search and data mining, ACM, pp 743–752

Wang D, Pedreschi D, Song C, Giannotti F, Barabási A-L (2011) Human mobility, social ties, and link prediction. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1100–1108

Xu-Rui G, Li W, Wei-Li W (2015) An algorithm for friendship prediction on location-based social networks. In: International conference on computational social networks, Springer, pp 193–204